BrainAlign: EEG-Vision Alignment via Frequency-Aware Temporal Encoder and Differentiable Cluster Assigner

Enze Shi, Huawen Hu, Qilong Yuan, Kui Zhao, Sigang Yu, and Shu Zhang✉

Center for Brain and Brain-Inspired Computing Research, School of Computer Science, Northwestern Polytechnical University, Xi'an, China
shu.zhang@nwpu.edu.cn

Abstract. While understanding visual processing in the human brain is fundamental for computational neuroscience, decoding objects from electroencephalography (EEG) remains challenging due to noisy neural dynamics during rapid image presentation and semantic misalignment in zero-shot settings. We propose BrainAlign, a novel framework leveraging contrastive learning to align EEG features with visual-language models (VLM). Our approach addresses three fundamental challenges: (1) We introduce a Frequency-Aware Temporal Encoder (FATE) using real Fast Fourier Transform with tunable bandpass filters to compress noisy signals while preserving temporal fidelity. (2) We develop a Differentiable Cluster Assigner (DCA) that dynamically optimizes channel grouping through cross-attention mechanisms, adaptively suppressing noise and enhancing task-relevant features. (3) We implement a self-supervised framework aligning EEG features with VLMs through contrastive learning. Extensive experiments demonstrate state-of-the-art performance on large-scale datasets, improving zero-shot retrieval accuracy by 5.85% and classification by 3.3%. Our work establishes new possibilities for brain-computer interfaces.

Keywords: Electroencephalography (EEG) · Contrastive learning · Dynamic channel clustering · Semantic alignment.

1 Introduction

The human brain recognizes objects from complex visual scenes within milliseconds of exposure [7], a capability reflected in the transient neural dynamics captured through EEG during visual stimulation [8,14]. A critical neuroscience question emerges: *Can we decode semantic information about perceived objects directly from EEG?* Three challenges persist: modeling temporal dynamics under rapid stimuli, capturing complex inter-channel dependencies, and aligning neural representations with visual semantics for zero-shot generalization.

Brain-Vision Semantic Alignment. Visual-language models (VLMs) with strong capabilities [17,4,15] offer unprecedented opportunities for bridging neural signals and semantic representations. While fMRI studies [19,9] have demon-

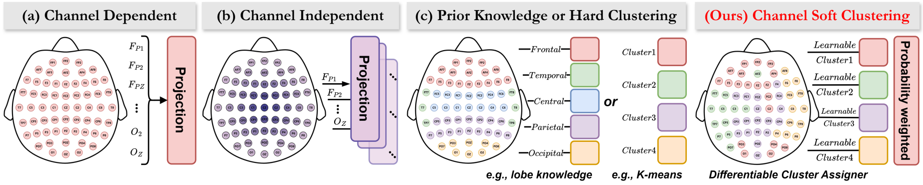

Fig. 1. EEG channel modeling strategy.

strated visual experience reconstruction using VLM-enhanced frameworks, EEG-based approaches face unique challenges due to low signal-to-noise ratios and temporal sparsity [22,14]. BrainAlign establishes a dynamic mapping between EEG features and VLM embeddings through contrastive learning, enabling knowledge transfer from visual-language domains to neural signal interpretation.

Neural Dynamics Modeling. Traditional approaches under rapid visual stimulation either struggle with spectral decomposition (time-domain methods [29,1]) or sacrifice temporal resolution (frequency-domain approaches [20]). Our FATE (Frequency-Aware Temporal Encoder) module introduces a hybrid architecture performing real FFT-based [2] spectral interpolation while maintaining temporal coherence through flexible bandpass filters and inverse mapping, preserving critical phase information often lost in conventional spectral methods.

Reinventing Multi-Channel Relationships. EEG channel interactions present complex modeling challenges. As shown in Fig. 1, current strategies either oversimplify dependencies (channel-independent models [16]), enforce rigid anatomical priors (brain-region partitioning [26]), or rely on static clustering [28]. Our DCA (Differentiable Cluster Assigner) module revolutionizes this paradigm through: (1) data-driven cluster centers updated via cross-attention, (2) gradient-preserving optimization of cluster assignments, and (3) adaptive masking for noise suppression. This approach identifies optimal channel interactions based on stimulus content and individual neural characteristics.

Our principal contributions are threefold: (1) The BrainAlign framework establishes end-to-end contrastive alignment between EEG signals and visual-language models, achieving 5.85% and 3.3% accuracy improvement in zero-shot object recognition. **(2)** The FATE module introduces a theoretically-grounded hybrid time-frequency architecture. By integrating real FFT-based spectral interpolation with learnable frequency projections and inverse temporal mapping, FATE creates more discriminative neural representations. **(3)** The DCA module fundamentally transforms EEG channel analysis through a novel end-to-end optimization framework for dynamic channel clustering.

2 Method

2.1 BrainAlign Architecture

As illustrated in Fig. 2, BrainAlign operates in two stages: EEG-VLM alignment and zero-shot inference. Given EEG trials $X_i \in \mathbb{R}^{C\times T}$ evoked by natural image

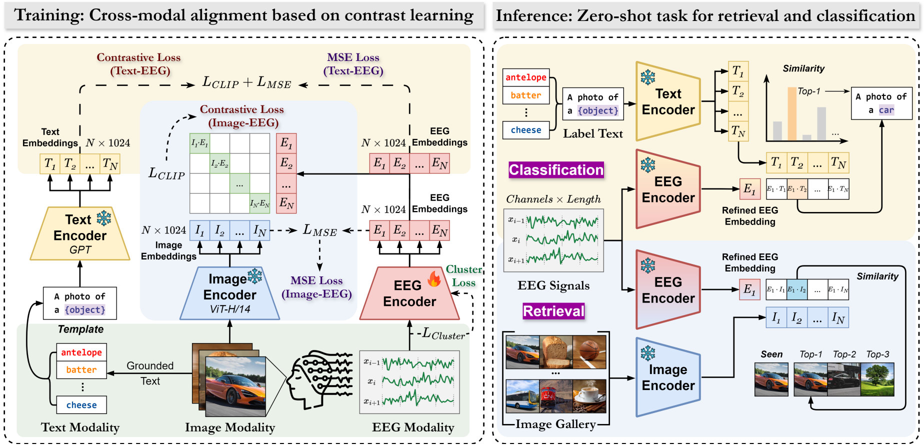

Fig. 2. Overview of the BrainAlign architecture and workflow (Training and Inference).

stimuli I_i, our EEG encoder first extracts latent features $e_i = f_{EEG}(X_i) \in \mathbb{R}^d$. Simultaneously, the paired stimulus image is encoded by a frozen VLM (i.e., OpenCLIP [17]) into semantic embeddings $v_i = f_{VLM}(I_i) \in \mathbb{R}^d$. The alignment process is formulated as a contrastive learning task and a regression task:

$$\mathcal{L}_{regress} = \frac{1}{N} \sum_{i=1}^{N} \|e_i - v_i\|_2^2 \tag{1}$$

$$\mathcal{L}_{contrast} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s(e_i, v_i)/\tau)}{\sum_{j=1}^{B} [\exp(s(e_i, v_j)/\tau) + \exp(s(e_j, v_i)/\tau)]} \tag{2}$$

where $s(\cdot)$ denotes cosine similarity, τ is a temperature parameter, and B is the batch size. The regression term $\mathcal{L}_{regress}$ employs mean squared error (MSE) to reconstruct target image features v_i from EEG signals x_i. Simultaneously, $\mathcal{L}_{contrast}$ encourages EEG representations to be similar to their corresponding visual stimuli while being dissimilar to other images in the batch. This dual-objective ensures the model preserves both modality-specific fidelity and cross-modal consistency. The classification task follows the same pipeline.

During inference, we perform zero-shot object recognition by: (1) encoding a test EEG signal using f_{EEG}, (2) computing its similarity with a gallery of VLM features extracted from candidate images (for retrieval) or with class prototypes derived from text prompts like "a photo of a [class]" (for classification), and (3) selecting the highest similarity match.

2.2 Frequency-Aware Temporal Encoder (FATE)

FATE Pipeline. As illustrated in Fig. 3-(a), FATE functions through a three-stage process: frequency transformation, spectral processing, and inverse mapping. Initially, we convert the time-domain EEG signal to the frequency domain

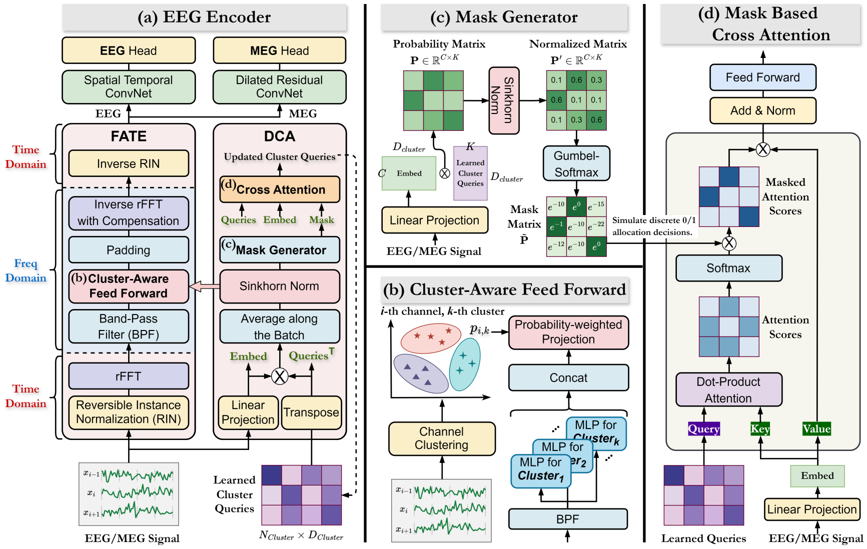

Fig. 3. Detailed structural presentation of BrainAlign's key components.

using the real Fast Fourier Transform (rFFT) [2]: $\hat{X} = \text{rFFT}(X) \in \mathbb{C}^{C \times F}$, where F represents the number of frequency components.

Additionally, FATE incorporates bandpass filtering to optimize compatibility with EEG signals. The bandpass filter effectively eliminates extraneous frequency components outside the relevant bands, compressing the model representation while preserving essential EEG characteristics. Subsequently, a cluster-aware multilayer perceptron encodes the frequency-domain information. Although these operations are performed in the frequency domain, the third stage enables supervised training of time-domain features using MSE, employing the inverse real Fast Fourier Transform (irFFT) to map the processed frequency representation back to the time domain: $Z = \text{irFFT}(\hat{Z}) \in \mathbb{R}^{C \times T'}$, where T' may differ from the original signal length T depending on the spectral processing, necessitating zero-padding prior to the inverse transformation. This time-domain representation preserves the essential temporal patterns while eliminating noise and irrelevant frequencies.

Cluster-aware Feed Forward. Rather than projecting EEG channels individually or in mixture, or assigning specialized encoders to different brain regions or fixed channel combinations, we allocate a separate Feed Forward network to each soft cluster to capture underlying shared frequency patterns within the cluster [3]. As depicted in Fig. 3-(b), $h_{\theta_k}(\cdot)$ represents the linear layer for the k-th cluster with weights θ, and z_i denotes the hidden embedding of the i-th channel prior to projection. Consequently, the final frequency mapping constitutes the weighted average of outputs from all cluster feed-forward layers according to $p_{i,k}$, e.g., $Y_i = \sum_k p_{i,k} h_{\theta_k}(Z_i)$.

2.3 Differentiable Cluster Assigner (DCA)

Channel Clustering with Learned Queries. DCA adaptively learns channel groupings through a fully differentiable clustering mechanism, enabling end-to-end optimization within deep learning pipelines, as depicted in Fig. 3-(a). We initialize K learnable cluster embeddings, denoted as $\{c_1, ..., c_K\}$, where each $c_k \in \mathbb{R}^d$, with d representing the hidden dimension. Given an EEG input X, each channel is transformed into a d-dimensional embedding h_i using linear projection. To determine the association between channel i and cluster k, we compute the probability $p_{i,k}$ using the following equation:

$$p_{i,k} = \text{Normalize}(\frac{c_k^\top h_i}{\|c_k\|\|h_i\|}) \in [0,1] \tag{3}$$

Following the process illustrated in Fig. 3-(c), we employ a reparameterization technique [11] to derive the clustering mask matrix $\mathbf{M}$, where each element $\mathbf{M}_{ik}$ approximates a Bernoulli distribution. Higher probability values $p_{i,k}$ translate to $\mathbf{M}_{ik}$ values closer to 1, indicating strong association between the channel and cluster k. Subsequently, we utilize mask-based cross-attention to update learnable cluster queries, as shown in Fig. 3-(d). Defining $\mathbf{C} = [c_1, ..., c_K] \in \mathbb{R}^{K \times d}$ as the cluster embedding matrix and $\mathbf{H} = [h_1, \cdots, h_C] \in \mathbb{R}^{C \times d}$ as the channel embedding matrix, calculations proceed as follows:

$$\hat{\mathbf{C}} = \text{Normalize}\left(\exp(\frac{(W_Q\mathbf{C})(W_K\mathbf{H})^\intercal}{\sqrt{d}}) \odot \mathbf{M}^\intercal\right) W_V\mathbf{H} \tag{4}$$

Here, W_Q, W_K, and W_V represent learnable weight matrices.

Key Mechanism of Cluster Loss. To discover latent neurophysiological patterns, we introduce a spectral clustering-inspired regularization $\mathcal{L}_{cluster}$. Let $\mathbf{P} \in [0,1]^{C \times K}$ denote the raw cluster probability matrix. To enable gradient-based optimization while approximating discrete assignments, we apply Gumbel-Softmax relaxation:

$$\tilde{\mathbf{P}}_{ck} = \frac{\exp\left((\log \mathbf{P}_{ck} + G_{ck})/\gamma\right)}{\sum_{k'=1}^{K} \exp\left((\log \mathbf{P}_{ck'} + G_{ck'})/\gamma\right)}, \quad G_{ck} \sim \text{Gumbel}(0,1) \tag{5}$$

where γ is an annealing temperature. This produces a softened assignment matrix $\tilde{P}$, which is used for similarity-aware loss computation.

The cluster loss combines three terms:

$$\mathcal{L}_{cluster} = -\text{Tr}(\tilde{\mathbf{P}}^\top \mathbf{S}\tilde{\mathbf{P}}) + \text{Tr}\left((\mathbf{I} - \tilde{\mathbf{P}}\tilde{\mathbf{P}}^\top)\mathbf{S}\right) + \lambda \sum_{c,k} -\mathbf{P}_{ck} \log \mathbf{P}_{ck} \tag{6}$$

where Tr indicates a trace operator, $\mathbf{S}$ denotes the channel similarity matrix. The first term maximizes similarities within clusters, while the second penalizes similarities between different clusters. The entropy term prevents trivial solutions where all channels collapse into a single cluster. The overall loss function thereby becomes $\mathcal{L} = \alpha\mathcal{L}_{regress} + (1-\alpha)\mathcal{L}_{contrast} + \beta\mathcal{L}_{cluster}$, where α governs the trade-off between regression accuracy and cross-modal alignment, while β regulates the strength of structural regularization.

3 Experiments and Results

3.1 Datasets and Settings

We evaluate BrainAlign on the THINGS-EEG dataset [8], which comprises EEG recordings from 10 participants exposed to RSVP stimuli. The experiment utilized a time-optimized paradigm to minimize artifacts while maintaining participant engagement. The dataset contains 16,540 training samples (1,654 unique concepts × 10 images × 4 repetitions) and 16,000 test samples (200 concepts × 1 image × 80 repetitions), with stimuli presented in pseudo-randomized sequences to mitigate order effects.

The preprocessing method is consistent with the original paper of the dataset [8]: trials were segmented into 1,000 ms epochs aligned to stimulus onset, followed by baseline correction using the 200 ms pre-stimulus interval, we retained all channels and downsampled the data to 250 Hz. All training and testing procedures are based on Pytorch 2.1.2, running on NVIDIA RTX 4090 and CUDA12.4.

3.2 Overall Performance

Our proposed BrainAlign framework establishes new state-of-the-art performance across all evaluation protocols, demonstrating significant improvements over existing EEG decoding paradigms. As shown in Table 2 and Fig. 4-(a), BrainAlign achieves 30.55% Top-1 and 59.90% Top-5 accuracy in zero-shot EEG retrieval, surpassing the best baseline (ATM-S) by 5.85% and 4% respectively. Crucially, when integrating competing EEG encoders (NICE, ATM-S) into our framework, we observe consistent gains, validating our architecture's general superiority independent of encoder choice. Traditional EEG models like EEGNetV4 show competitive 2-Way accuracy (95.00%) but collapse in fine-grained tasks (69% 10-Way vs. our 76.05%), while foundation models (CBraMod: 19.65% Top-1) lag significantly despite pretraining. Our framework's cross-task robustness is further evidenced by 10.85% Top-1 classification accuracy. The architectural advantages generalize beyond EEG: on MEG data, BrainAlign's DRConv variant achieves 84.40% 2-Way accuracy.

Table 1. Retrieval performance (accuracy %) comparison using LOSO cross-validation.

Methods	**Top-1**	**Top-5**	**2-Way**	**4-Way**	**10-Way**
MLP	4.46±0.81	15.26±2.34	75.80	55.08	34.05
EEGNetV4 [13]	6.25±2.56	20.95±5.73	82.85	64.65	42.35
CBraMod (Finetune) [25]	6.60±2.21	20.30±5.20	80.25	61.45	42.55
FoME (Finetune) [21]	3.57±1.66	10.43±3.33	62.50	48.51	29.35
NICE (Our Framework)	8.70±2.38	26.10±4.52	84.50	67.35	49.10
ATM-S [14]	11.84±4.80	**33.73±8.73**	87.36	**72.80**	53.80
BrainAlign (Ours)	**12.40±3.31**	30.25±6.04	**88.50**	72.50	**59.13**

Table 2. Evaluation results (accuracy %) for zero-shot retrieval and classification tasks based on EEG/MEG datasets. The test set contains 200 classes and performance is evaluated using Top-1 and Top-5 accuracies as well as 2-way, 4-way and 10-way accuracies. We present a comprehensive comparison of different model types (EEG-specific models, EEG foundation models, time series models, and EEG-image models). The best result is highlighted in **bold**, and the second best is highlighted with underline.

Model Type	**Methods**	**Top-1**	**Top-5**	**2-Way**	**4-Way**	**10-Way**
	Chance level	0.50	2.50	50.00	25.00	10.00
Zero-shot Retrieval (EEG Dataset)						
EEG Model	EEG Conformer [23]	3.98	17.10	76.17	56.29	34.72
	ShallowFBCSPNet [18]	6.10	16.53	74.32	53.97	33.48
	MLP	10.50	33.00	83.50	62.50	41.50
	EEGNetV4 [13]	16.50	50.00	**95.00**	79.05	69.00
EEG Foundation Model	BrainBERT (Probe) [24]	5.00	2.50	52.00	25.80	11.20
	BrainBERT (Finetune) [24]	1.00	3.20	50.70	26.30	11.70
	Neuro-GPT (Probe) [5]	0.73	2.50	50.75	23.90	10.50
	Neuro-GPT (Finetune) [5]	8.26	25.49	71.60	51.50	31.50
	FoME (Probe) [21]	0.83	2.33	49.17	25.67	11.67
	FoME (Finetune) [21]	8.17	20.33	78.00	61.50	41.33
	LaBraM (Finetune) [12]	0.50	2.50	50.65	24.85	9.35
	CBraMod (Finetune) [25]	19.65	45.30	90.70	79.95	63.65
Time Series Model	PatchTST [16]	19.00	46.80	92.80	81.70	66.80
	Dlinear [27]	23.25	54.70	93.90	85.45	71.80
EEG-Image Model	NICE [22]	20.08	49.43	93.55	83.97	68.53
	NICE (Our Framework)	22.85	52.90	93.68	85.40	71.60
	ATM-S [14]	24.70	55.90	94.00	86.55	72.95
	ATM-S (Our Framework)	25.80	59.55	94.95	**88.50**	75.90
	BrainAlign (Ours)	**30.55**	**59.90**	**95.00**	88.10	**76.05**
Zero-shot Classification (EEG Dataset)						
Type	**Methods**	**Top-1**	**Top-5**	**2-Way**	**4-Way**	**10-Way**
EEG Model	MLP	2.80	9.70	69.30	47.70	25.40
	EEGNetV4 [13]	3.10	13.70	75.50	53.75	32.05
EEG Foundation Model	BrainBERT (Probe) [24]	0.60	2.70	49.30	24.90	9.90
	BrainBERT (Finetune) [24]	1.00	3.40	53.50	25.60	10.70
	Neuro-GPT (Probe) [5]	0.45	2.75	69.30	42.60	23.50
	Neuro-GPT (Finetune) [5]	1.72	9.25	49.80	24.50	9.60
	CBraMod (Finetune) [25]	5.95	16.35	74.95	55.20	33.80
EEG-Image Model	NICE (Our Framework)	7.25	26.60	83.40	66.85	46.65
	ATM-S [14]	7.55	22.60	82.75	65.40	43.25
	BrainAlign (Ours)	**10.85**	**28.05**	**84.70**	**69.55**	**49.25**
Zero-shot Retrieval (MEG Dataset [10])						
Type	**Methods**	**Top-1**	**Top-5**	**2-Way**	**4-Way**	**10-Way**
EEG-Image Model	BrainAlign (STConv)	5.55	16.85	79.25	58.05	37.5
	BrainAlign (DRConv [6])	**9.05**	**27.05**	**84.4**	**69.2**	**48.2**

BrainAlign also demonstrates superior cross-participant generalization in leave-one-subject-out (LOSO) retrieval, achieving 12.40% Top-1 accuracy (a 4.7% improvement over standard ATM-S), as shown in Table 1. The code will be released at https://github.com/1061413241/BrainAlign.

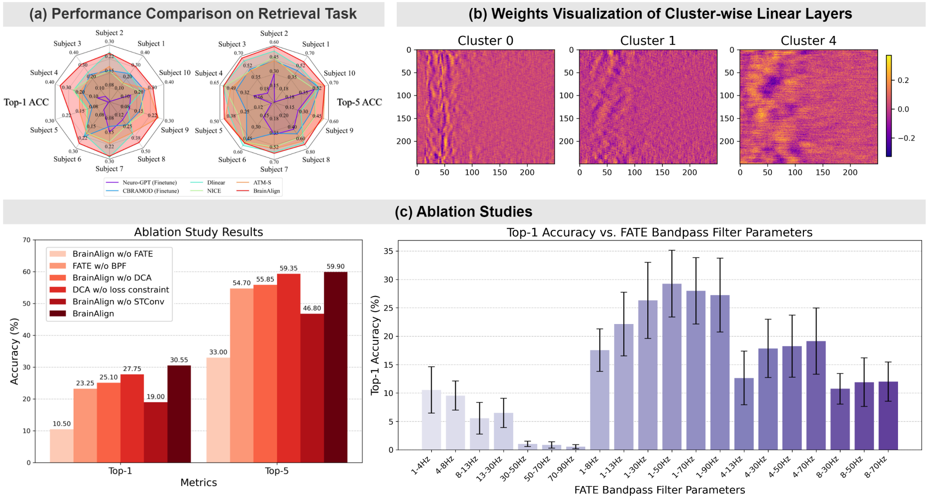

Fig. 4. Overview of model interpretation and ablation studies.

3.3 Model Interpretation and Ablation Study

Fig. 4-(b) visualizes the activation states of feedforward layers corresponding to distinct clusters, revealing diverse patterns that guide the model's feature learning. For instance, Cluster-0 shows concentrated activation in the left region, potentially focusing on capturing localized periodic variations, while Cluster-4 exhibits a more widespread activation pattern, suggesting an emphasis on global feature extraction.

As shown in Fig. 4-(c), BrainAlign achieves peak 30.55% Top-1 accuracy when all components are active, with systematic performance degradation observed under ablation conditions. Parameter sensitivity analysis reveals optimal performance when FATE's bandpass filter operates in the 1-50 Hz range ($\theta/\alpha/\beta$ waves). This aligns with neuroscientific evidence linking semantic processing to low/mid-frequency oscillations.

4 Conclusion

We propose BrainAlign, a novel framework for decoding visual semantics from EEG signals using contrastive learning to bridge neural and visual-language representations. Our work advances the state-of-the-art in EEG-based object recognition through three key innovations: (1) the FATE module uniquely preserves both spectral and temporal fidelity in millisecond-level visual processing, (2) the DCA module revolutionizes channel relationship modeling through gradient-driven dynamic clustering, and (3) a robust contrastive learning strategy that enables zero-shot generalization with significant performance gains (5.85% in retrieval, 3.3% in classification). While BrainAlign demonstrates unprecedented capabilities in neural decoding, limitations remain in handling extreme noise conditions and generalizing across diverse recording environments. Future work will focus on extending our framework to continuous EEG decoding during natural viewing conditions, incorporating multimodal fusion with other neuroimaging techniques to enhance our understanding of brain-vision relationships.

Acknowledgments. This study was funded by National Natural Science Foundation of China (62376219 and 62006194), Faculty Construction Project (Grant No. 25SH02010044), and Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Altaheri, H., Muhammad, G., Alsulaiman, M.: Physics-informed attention temporal convolutional network for eeg-based motor imagery classification. IEEE transactions on industrial informatics **19**(2), 2249–2258 (2022)
2. Brigham, E.O., Morrow, R.E.: The fast fourier transform. IEEE spectrum **4**(12), 63–70 (1967)
3. Chen, J., Lenssen, J.E., Feng, A., Hu, W., Fey, M., Tassiulas, L., Leskovec, J., Ying, R.: From similarity to superiority: Channel clustering for time series forecasting. Advances in Neural Information Processing Systems **37**, 130635–130663 (2025)
4. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
5. Cui, W., Jeong, W., Thölke, P., Medani, T., Jerbi, K., Joshi, A.A., Leahy, R.M.: Neuro-gpt: Towards a foundation model for eeg. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
6. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., King, J.R.: Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence **5**(10), 1097–1107 (2023)
7. DiCarlo, J.J., Cox, D.D.: Untangling invariant object recognition. Trends in cognitive sciences **11**(8), 333–341 (2007)
8. Gifford, A.T., Dwivedi, K., Roig, G., Cichy, R.M.: A large and rich eeg dataset for modeling human visual object recognition. NeuroImage **264**, 119754 (2022)
9. Gong, Z., Bao, G., Zhang, Q., Wan, Z., Miao, D., Wang, S., Zhu, L., Wang, C., Xu, R., Hu, L., et al.: Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. Advances in Neural Information Processing Systems **37**, 51655–51683 (2025)

10. Hebart, M.N., Contier, O., Teichmann, L., Rockter, A.H., Zheng, C.Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., Baker, C.I.: Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. Elife **12**, e82580 (2023)
11. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)
12. Jiang, W., Zhao, L., Lu, B.l.: Large brain model for learning generic representations with tremendous eeg data in bci. In: The Twelfth International Conference on Learning Representations
13. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)
14. Li, D., Wei, C., Li, S., Zou, J., Qin, H., Liu, Q.: Visual decoding and reconstruction via eeg embeddings with guided diffusion. arXiv preprint arXiv:2403.07721 (2024)
15. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
16. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: The Eleventh International Conference on Learning Representations
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
18. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for eeg decoding and visualization. Human brain mapping **38**(11), 5391–5420 (2017)
19. Scotti, P.S., Tripathy, M., Villanueva, C.K.T., Kneeland, R., Chen, T., Narang, A., Santhirasegaran, C., Xu, J., Naselaris, T., Norman, K.A., et al.: Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. arXiv preprint arXiv:2403.11207 (2024)
20. Shi, E., Yu, S., Kang, Y., Wu, J., Zhao, L., Zhu, D., Lv, J., Liu, T., Hu, X., Zhang, S.: Meet: A multi-band eeg transformer for brain states decoding. IEEE Transactions on Biomedical Engineering **71**(5), 1442–1453 (2023)
21. Shi, E., Zhao, K., Yuan, Q., Wang, J., Hu, H., Yu, S., Zhang, S.: Fome: A foundation model for eeg using adaptive temporal-lateral attention scaling. arXiv preprint arXiv:2409.12454 (2024)
22. Song, Y., Liu, B., Li, X., Shi, N., Wang, Y., Gao, X.: Decoding natural images from eeg for object recognition. In: The Twelfth International Conference on Learning Representations
23. Song, Y., Zheng, Q., Liu, B., Gao, X.: Eeg conformer: Convolutional transformer for eeg decoding and visualization. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 710–719 (2022)
24. Wang, C., Subramaniam, V., Yaari, A., Kreiman, G., Katz, B., Cases, I., Barbu, A.: Brainbert: Self-supervised representation learning for intracranial electrodes. In: International Conference on Learning Representations. ICLR (2023)
25. Wang, J., Zhao, S., Luo, Z., Zhou, Y., Jiang, H., Li, S., Li, T., Pan, G.: Cbramod: A criss-cross brain foundation model for eeg decoding. arXiv preprint arXiv:2412.07236 (2024)

26. Yi, K., Wang, Y., Ren, K., Li, D.: Learning topology-agnostic eeg representations with geometry-aware modeling. Advances in Neural Information Processing Systems **36**, 53875–53891 (2023)
27. Zeng, A., Chen, M., Zhang, L., Xu, Q.: Are transformers effective for time series forecasting? In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 11121–11128 (2023)
28. Zhang, C., Sun, L., Cong, F., Kujala, T., Ristaniemi, T., Parviainen, T.: Optimal imaging of multi-channel eeg features based on a novel clustering technique for driver fatigue detection. Biomedical Signal Processing and Control **62**, 102103 (2020)
29. Zhang, S., Shi, E., Wu, L., Wang, R., Yu, S., Liu, Z., Xu, S., Liu, T., Zhao, S.: Differentiating brain states via multi-clip random fragment strategy-based interactive bidirectional recurrent neural network. Neural Networks **165**, 1035–1049 (2023)